Deep Learning Practical Work 2-b
Visualizing Neural Networks

Caterina Leonelli & Luisa Neubauer

December 29, 2023

1 Introduction

This document contains a list of questions and their corresponding answers.

2 Questions and Answers

1. **Question 1:** Show and interpret the results
    - *Answer:* The saliency map in general conforms to our guess of where one would look in order to identify the object in an image. For example in the case of the class 'daisy' (see figure 3, we have high saliency at the disk flowers surrounded by elevated activation in the region of the petals and low to none in the background. Interestingly, the pixels that contribute to the class score are highly localized for some classes (e.g. Cardigan Welsh, hay, ...) whereas for other, such as 'modem' or 'spatula' , the contributing pixel region is spread out or covers almost the entire image. In this case, the network might have learned to identifiy assosciated objects that then by induction help to predict the desired class. We hypothesize that for example in the case of 'spatula' the network's filter also activate for 'pancakes' or 'pan' because these objects typically are co-occuring.
2. **Question 2:** Limits of this technique of visualizing the impact of different pixels
    - *Answer:* This technique is good at showing the impact of a single pixel but doesn't make any statement about interactions between pixels (pairwise influence). Also, different models can have widely different saliency maps for the same image.
3. **Question 3:** Can this technique be used for a different purpose than interpreting the network?

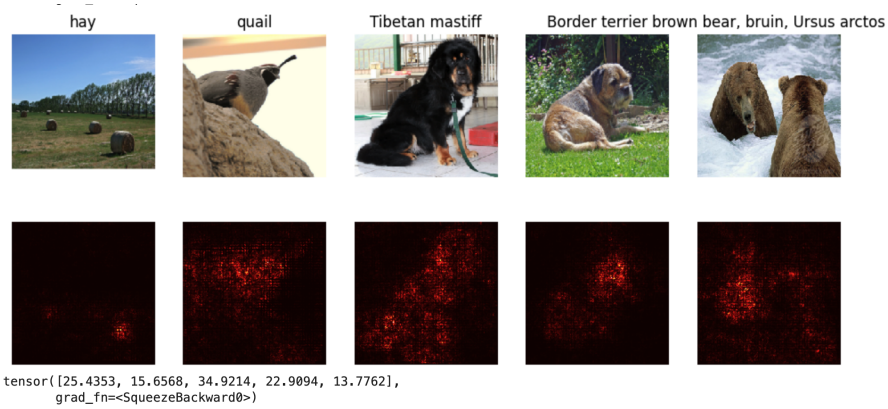

Figure 1: Saliency map samples for several categories

- *Answer:* One idea how to repurpose the saliency map would be for object detection. As the saliency activation is in most cases limited to the region where we - as humans - localize the object within the frame, we could use the saliency pixels to find the bounding boxes for our target objects.

4. **Question 4:** (Bonus) - Test with a different network, for example VGG16

   - *Answer:* We employ a pretrained vision transformer and compare the saliency maps. Interestingly, the regions of activation are very different across different architectures. For VIT, the acitvated regions are much more concentrated. Additionally, we see artifacts of the VIT's architectures.

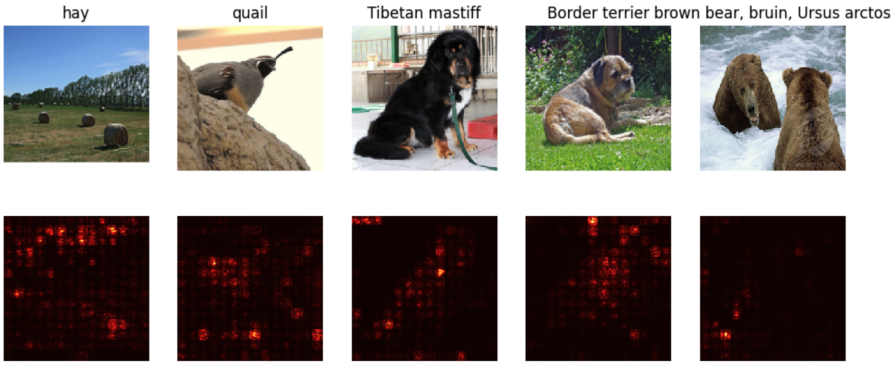

Figure 2: Saliency map samples for several categories using VIT

5. **Question 5:** Show and interpret the obtained results

- *Answer:* Although the transformed image looks almost unchanged to the human eye and is still being identified by a human as 'quail', the network has 'mislabeled' it as a stingray due to the marginal differences. The CNN is highly sensitive to small changes in pixel intensity.

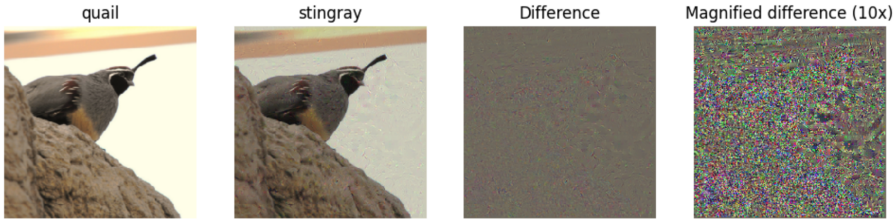

Figure 3: Magnified saliency map

6. **Question 6:** Consequences of this method on CCNs

   - *Answer:* This way, we can test how robust the CNN representations of our categories are with respect to e.g. input noise. Also, they can be used to find flaws in the network or to find weaknesses/security risks.

7. **Question 7:** (Bonus) - Discuss the limits of this naive way to construct adversarial images. Can you propose some alternative or modified ways? (You can base these on recent research).

   - *Answer:* First, the images generated this way do not give us any semantic information of why this could be classified as say a stingray instead of a quail. The individual pixel differences do not form a solid explanation of the CNN behaviour. Also, they might be highly sensitive to the model itself. The adversarial example generated this way, might or might not be an adversarial example for another network. One way to use these adversarial examples would be to feed them to the network in a separate training process with the correct labels to improve the model's robustness. One could also try to find the decision boundary of the NN in a systematic manner, sample from there and create adversarial examples from this regions.

8. **Question 8:** Show and interpret obtained results (class visualization)

   - *Answer:* The image shown in figure 4 is the result of the class visualization technique for the category 'Gorilla'. What this technique does is, it aims at learning a stereotypical (one that maximized a certain category) representation of a category. The pattern in the image shows features and shapes associated with gorillas such as the

head or the characteristic eyes. The fact that these somewhat abstract features appear in clusters indicates that the CNN expects the features to feature within a certain part of the input images.

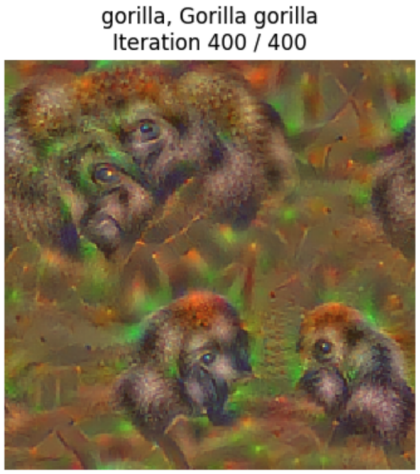

Figure 4: Class visualization gorilla

9. **Question 9:** Adjust training parameters for class visualization

   - *Answer:* We have found the combination of parameters $N_{ITER} = 500$ $\mathrm{L2}_{REG} = 1e-2$ to be ideal for the class visualization. More training iterations did not yield more expressive class visualizations (as judged by the human eye).

$$\mathrm{L2}_{reg} = 1e-3 \tag{1}$$
$$\mathrm{learning}_{rate} = 5, \mathrm{num}_{\mathrm{iterations}} = 200 \tag{2}$$
$$\mathrm{blur}_{every} = 10, \mathrm{max}_{\mathrm{jitter}} = 16 \tag{3}$$

10. **Question 10:** Class visualization on ImageNet image (Try to use an image from ImageNet as the source image instead of a random image (parameter init_{img}). You can use the real class as the target class. Comment on the interest of doing this)

    - *Answer:* We have started out with a 'normal' picture of a daisy that is within the ImageNet dataset and applied the transformations for the category daisy. The resulting image can be found in figure **??**. This way, we made the image according to the CNN extremely 'daisy'. The network added features that it found lacking such as the yellow calyx.

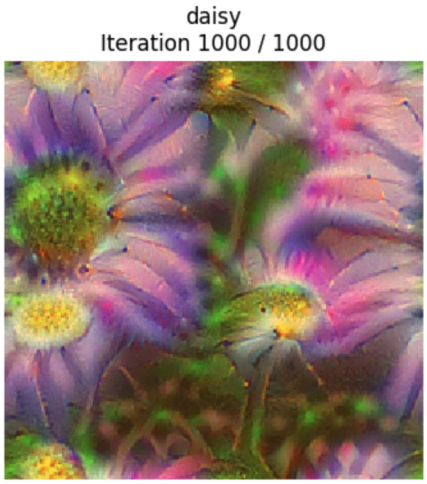

Figure 5: Class visualization on class 'daisy' with starting image depicting a daisy

11. **Question 11:** Bonus - Test with another network.

    - *Answer:* We reiterated the class visualization procedure on VGG16. Interestingly, despite a big resemblance with the picture obtained this way on our first CNN, the way VGG16 imagines a snail differs slightly. For example, VGG16 clearly pictures some snail shells whereas for the CNN, the antlers and eyes seem to be important.

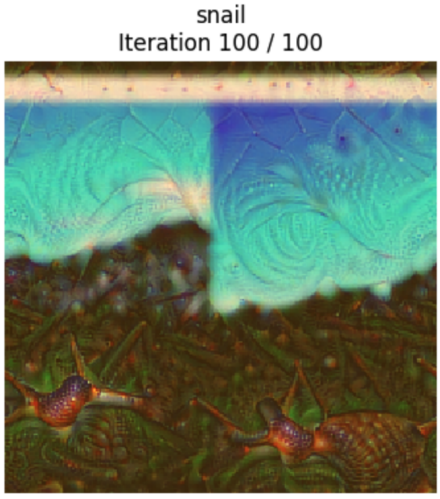

Figure 6: Class visualisation snail (VGG16)

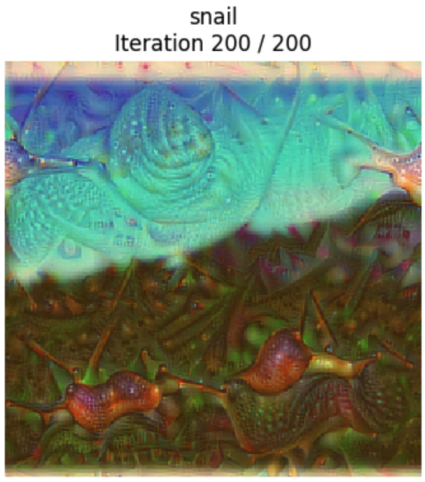

Figure 7: Class visualisation snail (VGG16) after 200 iters